Interjudge Reliability in the Measurement of Pitch Matching

A Senior Honors Thesis

Presented in partial fulfillment of the requirements for graduation with distinction in Speech and Hearing Science in the Undergraduate College of The Ohio State University

By

Stephanie L. Joseph

The Ohio State University
June, 2007

Project Advisor: Dr. Michael Trudeau and Dr. Jan Weisenberger, Department of Speech and Hearing Science

Abstract

In the clinical analysis and treatment of voice disorders, software packages allow the extraction of acoustic voice parameters, such as fundamental frequency, from patient vocal productions. The utility of such software packages in the clinical situation, however, is dependent upon consistency in their use and in interpretation of the analyses they provide. The present study focused on the extraction and interpretation of voice pitch data. Samples from eight trained singers who were asked to match the pitch of both pure tones and synthesized human voice were submitted to the Kay Elemetrics Computerized Speech Lab for analysis. Estimates of voice pitch were made by three judges, and the inter-rater reliability was measured. Results indicated that judges were very consistent in the estimation of token duration but varied widely in estimates of fundamental frequency. These results have important implications for use of such software packages in clinical diagnosis and treatment.

Acknowledgments

I would like to thank Dr. Michael Trudeau for giving me the opportunity to work with him and allowing me to be “a fly on the wall” in his office for hours on end, where I absorbed a great deal of information while also gathering data at his computer. I would like to thank Dr. Jan Weisenberger who has dedicated so much of her time, not only in helping me to write my thesis, but also in calming me down and keeping me sane during the many long hours spent in her office. I would also like to thank Stephanie Finical and Allison Masty who worked with me to gather the data and who made this experience fun.

The present study was supported by the ASC Undergraduate Research Scholarship and the SBS Undergraduate Research Scholarship.

Table of Contents

Chapter One: Introduction and Literature Review

Accurately matching the pitch of a sound is an important skill for trained singers. It is a complicated process in which the singer must process the stimulus, organize and initiate a response and then produce the response so that it corresponds in fundamental frequency to the stimulus. The fundamental frequency of the voice must match the pitch of the fundamental frequency produced by the instrument or voice token. The production of the sound involves control and coordination of the respiratory, phonation, resonance and afferent systems. This coordination and control improves with training and practice.

Studies of Pitch Matching Performance:

Studies have shown that trained singers accurately match pitch better than non-trained singers; however, there is a natural distribution of ability within the population. Thomas Murry (1990) tested differences in pitch matching between singers and nonsingers using a fundamental frequency matching test. He tested laryngeal accuracy and the effects of training on five female trained singers versus five females with no singing experience. The participants in each group were presented with a tone through a loudspeaker and asked to match the pitch. Only the first five waveforms of the responses were analyzed to minimize the adjustments of the singers based on their auditory abilities. The results showed that singers were more accurate in matching the fundamental frequency than the nonsingers. The singers were also less variable, performing with the same accuracy in the first waveform as compared to the mean of the first five. Murry (1990) concluded that training and experience, as well as skill, explain the singers' increased abilities in

accurately matching fundamental frequency to a target. Leonard et al. (1988) also conducted an experiment comparing singers to nonsingers. This study, however, focused on time, speed and damping characteristics of singers and nonsingers performing a vocal shadowing task. Five male trained singers and five males with no professional singing experience listened to tones through earphones and sang the note into a microphone. The tones were square-wave modulated changes of triangular wave pulse-trains. The results showed that the singers generally needed less time to reach the pitch changes than did the nonsingers when the pitch was raised as well as lowered. Their faster speeds were attained by moving directly to the targets. Both groups of subjects, however, employed four different patterns of achieving pitches, including hit, overshoot, undershoot and oscillate. Leonard et al. (1988) suggested that it is possible that the singers learned to use the damping patterns differently and more successfully than did the nonsingers.

Stimulus Factors Affecting Pitch Matching:

Within the general population, the significance of pitch matching plays different roles. An article by Small and McCachern (1983) discusses the implications in teaching first grade children how to sing. The teaching is often done using a live model, and the study looked at the differences the children experienced to see if they could match pitch more accurately with a female than with a male model. Fifty-five first grade students in a public school were taught to sing listening to a recording of either a male or female voice during five, thirty-minute sessions. Results indicated that some of the students had no difficulty whether the teacher was a male or a female, and others had difficulty with both models. This result suggests that a degree of skill or talent affects pitch matching to a greater degree than the gender of the eliciting stimulus.

In the mechanisms that underlie the ability to match a pitch, questions arise as to whether certain stimulus factors facilitate pitch matching ability. Yarbrough et al. (1992) conducted a study that looked at the effect of vibrato on pitch matching accuracy between certain and uncertain singers. The subjects were students from kindergarten through third grade and were classified as certain or uncertain singers based on their pitch matching ability before the task began. Each student responded to a child model, a female non-vibrato singing accurately 100% of the time, and a female vibrato singing accurately 79.53% of the time. The results indicated that the vibrato did have an effect on the pitch matching ability of the uncertain singers, eliciting significantly fewer correct responses from them. The certain singers sang the series of pitches correctly, whether or not the vibrato deviated slightly; however, the uncertain singers matched the vibrato model whether she was singing the descending minor third correctly or not. Yarbrough et al. found that responses to the non-vibrato model were similar across grade level and gender, suggesting that this model would elicit more consistent and accurate pitch matching.

A study by Tervaniemi (2000) focused on the perceptual aspect of pitch matching. He measured auditory evoked responses to pure tones and spectrally rich stimuli and found that observers were more sensitive to spectrally rich stimuli. The spectrally rich stimuli carry more spectral and temporal information than pure tones. Tervaniemi speculates that because the human auditory system has extensive experience with spectrally rich stimuli such as speech and music, it is possible that it reacts more vigorously to changes in spectrally complex rather than simple, ecologically unrepresentative information. The implication is that spectrally complex stimuli would also elicit more accurate pitch productions.

Ives (2002) directly tested whether this perceptual discrimination ability was reflected in actual pitch matching performance. He studied how well singers could match the pitch of a

human synthesized vocal model compared to a pure tone stimulus. Of specific interest was the minimum length of time required for the human auditory system to process pitch information in order to reproduce it accurately. In this study, the participants included ten male students who were trained singers in The Ohio State University Men's Glee Club. Their responses to stimulus tokens ranging from 130 to 311 Hz fundamental frequency were recorded and analyzed using the computer program, Cool Edit 2000. The study compared the fundamental frequency of the first identifiable period of the vocalization to the average response measured in the first five periods, in order to determine if there was a difference between the two measures in the pitch matching accuracy.

Surprisingly, Ives did not find a difference in pitch matching accuracy to the pure tones versus the human synthesized vocal model, which is a more spectrally rich sound. However, this study did yield other interesting findings. He did find a significant difference between the average of the first period and the average of the first five periods, showing that a period of adjustment had taken place.

Follow-up work by Ameer (2001) investigated the time required to achieve phonatory stability by the same singers used in Ives' study. She found that on average, six cycles were required to achieve a stable fundamental frequency. A study by Curran (2004) looked more closely at Ives' measurement techniques. Ives had examined only the first five periods of the response in measuring pitch matching accuracy. Curran's analysis used the entire waveform in determining pitch matching accuracy. She found that the measurement technique of using the entire response provided the most information and therefore best represented the pitch matching accuracy.

Methodological Issues and User Variability:

Several additional methodological details in Ives' study warrant further examination. One issue deals with the software program used in his analysis. Cool Edit 2000 has certain limitations. One example of these limitations is in the fact that it uses zero crossings to analyze the fundamental frequency. This becomes a problem with complex waveforms, such as the human voice, which may have two zero crossings in one cycle of the fundamental frequency. Algorithms for pitch extraction are used in every software program and it is possible that different software systems exploring different pitch extraction algorithms could derive very different estimates of voice pitch. This idea was investigated by Smits et al. (2005) in a different comparison of two popular software programs, Dr. Speech (DRS) and Computerized Speech Lab (CSL). The study compared the acoustic measurements performed with DRS and CSL to determine differences and similarities between the systems. The parameters studied included fundamental frequency, fundamental frequency standard deviation, absolute and relative jitter, relative shimmer and harmonics-to-noise ratio. The results showed that some of the parameters, such as fundamental frequency, were comparable, while others, such as relative shimmer and HNR, were comparable only with a transformation key, and still others, such as jitter and fundamental frequency standard deviation, were not comparable at all. The authors concluded that for jitter and shimmer data, there is generally a weak correlation due to the different algorithms used by the software programs to extract these parameters.

Methodological issues in Ives' study dealt with the fact that in the Cool Edit, the pitch extraction is performed on a specific segment of the waveform, which is highlighted by the program user. This pitch extraction process consists of a potential source of error derived from two sources. First is a simple variability in the accuracy of the human user in choosing the

waveform segment to highlight. This could be dependent on motor skills and previous practice. According to Slifkin and Newell (1998), the characteristics of performance are never exactly the same, even under the same task conditions and when the user is a practiced expert. There may also be a sensory basis when highlighting the waveform and the limits may vary depending on where the user perceives that the beginning of the waveform is located. The second possible basis for error occurs in situations where the waveform boundaries are ambiguous. Here it is possible that users will employ different strategies for marking sample boundaries. This may be particularly true for software programs like CSL that provide multiple options for resolving ambiguity.

To date, the possible effects of variability in performance of users of voice analysis software programs on extracted pitch of waveform samples have not been evaluated. In some situations, where only a broad estimation of pitch matching accuracy is needed, this might not be of concern. However, for highly skilled performers, such as trained singers, even small differences in extracted pitch might prove to be important. This could also be an issue in clinical usage of these software packages for voice patients, because objective acoustic measurements in voice clinics have become a substantial aspect of voice assessment in the last few decades (Smits et al., 2005).

The present study addressed the degree to which variability across users of the Kay Elemetrics Computerized Speech Lab (CSL) affects the obtained pitch values extracted from the vocal waveforms of trained singers. Waveform samples of recordings from Ives' original study were reanalyzed using the CSL. This re-analysis took advantage of the capabilities of a more advanced computer program that measures different parameters of the vocalizations. The focus of the present study was to explore the role of inter-evaluator differences in the determination of

pitch matching accuracy. Three different evaluators analyzed two thirds of the data from the previous studies, such that each token was re-analyzed by two different observers. If the reliability between the evaluators proved to be low, the conclusions of previous pitch matching studies with a single evaluator would need to be re-examined. At the present time, no computer program is able to determine the pitch matching accuracy of individual tokens without the intervention of a human evaluator. When the data are measured by human evaluators, the reliability of these evaluators becomes crucial. The results of this study should provide a reliable methodology for the evaluation of pitch matching accuracy that can form the basis of future studies in this area.

Chapter Two: Methodology

Test subjects and recording procedures for this study were obtained from a study by Ives (2002). The initial data for both of the studies are the same; however the analysis and questions that were raised were different.

Ives Methodology Summary

Subjects:

The test subjects for this study consisted of ten male students with a background in vocal music. This background consisted of a minimum of four years in formal voice instruction and six years in choral singing. The subjects reported no history of laryngeal pathology or voice disorders requiring phonosurgery or voice therapy. Hearing screenings were conducted on all subjects and confirmed that the subjects were within the normal threshold of 20 dB HL at the tested frequencies. The subjects warmed up at least two hours prior to testing for a minimum of ten minutes. They were seated in a sound-treated room and instructed to listen to a tone and match its pitch as quickly and accurately as possible. The subjects were given five trial pitches at durations ranging from 50 ms to 300 ms in order to familiarize them with the experimental task. The pitches used were within the range of those in the experimental task; however they were a different set of frequencies. Subjects had the opportunity to adjust the amplitude during the trial period so that no adjustments were necessary during the experiment.

Tokens:

There were eight target frequencies presented randomly to the subjects to avoid order effects. The frequencies ranged from 130.1 to 311.1 Hz (C to D#) and consisted of human synthesized voice or pure tones. The four token durations were set at 50 ms, 100 ms, 200 ms and 300 ms. There were five randomized token sets, each equally representing both token types, the four durations and eight frequencies. A total of 64 tokens appeared in each set, and the order in which the five lists were presented to each subject, was randomly determined.

Cool Edit 2000 was used to generate the pure tones while the male human synthesized voice tokens were generated using Sensimetrics, High-Level Parameter Speech Synthesis System version 2.2. The first three formants for the synthesized voice were 554 Hz, 916 Hz, and 2466 Hz, and a vibrato rate of 10 cycles per second was synthesized to have the voice tokens more closely simulate a human singing voice.

Recording Procedures:

The subjects were seated in a sound-treated room and told to listen for a tone from a speaker and match its pitch as quickly and as accurately as possible. After the training period, the tokens of either a synthesized voice or pure tones were played. The subjects were instructed to respond quickly, singing as loudly or softly as they chose. They were also informed that they could use either a “straight tone” or vibrato when they responded as long as they did not exceed two seconds. Two windows were open in Cool Edit 2000, one to play the sound file, and one to record both the token and the response. A custom made mixer/splitter allowed the sound token and pitch-matching attempts to be mixed and simultaneously recorded at 44.1 kHz sampling rate with a depth resolution of 16. This resulted in 2560 samples of pitch matching.

Analysis Procedures:

Cool Edit 2000 was used to obtain the duration of the first period and the first five periods and the response latency. The target stimuli and the sustained vowels were verified for each subject. The pitch matching accuracy of each subject's initial production was measured. Curran (2004), using Ives' data and Cool Edit 2000, analyzed the entire response to determine the overall pitch matching accuracy.

Methodology for Present Study

Analysis Procedures:

The program used to analyze the acoustic data from eight of the ten singers collected by Ives (2002) was Kay Computerized Speech Lab (CSL Model 4500). In previous studies, the data were analyzed using Cool Edit 2000. In the present study, three judges worked together, each analyzing two thirds of the data so that every token was evaluated twice. This study was designed to assess the interjudge reliability.

Before beginning the analysis of the data, the CSL Pitch Contour Analysis Program was set to a maximum analysis range of 0-500 Hz, as well as a maximum display range from 70 Hz to 500 Hz. Based on the previous analyses and tokens, it was determined that the pitch contour could exceed 350 Hz, which is the default setting. Thus, the maximum analysis range of 0-500 Hz would adequately cover the pitch contrast. Ives' files were then opened and the first response was located by listening to the recording and viewing the acoustic envelope. At the beginning of each file, the subject stated his name and the number of the file. Responses were then recorded. The response of the subject was easily identified on the screen because it had a much greater amplitude and duration than the eliciting tokens. The cursor was placed at the beginning of the

response and dragged across so that the response was highlighted. At this point only the selected data were viewed in the screen and the data were again selected with greater temporal resolution. The new selection was viewed and the process occurred a third time. At this point, only the response of the subject was viewed in the window. The *pitch contour* function was selected and a graph of the contour appeared in the bottom window. The information regarding the pitch contour was obtained by using the *information* function. Sample duration and average fundamental frequency were entered onto a Microsoft Excel spreadsheet.

Because a new program was being used to examine data that already existed, there were some difficulties. In a number of responses, the signal was not strong enough to be detected visually on the CSL program. In this case, the signal was played audibly using the *speak* function to assist in identifying the token which was highlighted, and then the pitch contour was obtained. The graph of the pitch contour reflected the response in the window above, whether or not it could be detected visually. The response could then be selected in the top window by linking the two windows and selecting the response in the bottom window.

Chapter Three: Results

Figures 1-3 show extracted duration values for pairs of judges as a function of stimulus fundamental frequency. As can be seen, there is a great deal of agreement in duration estimates for all three judges.

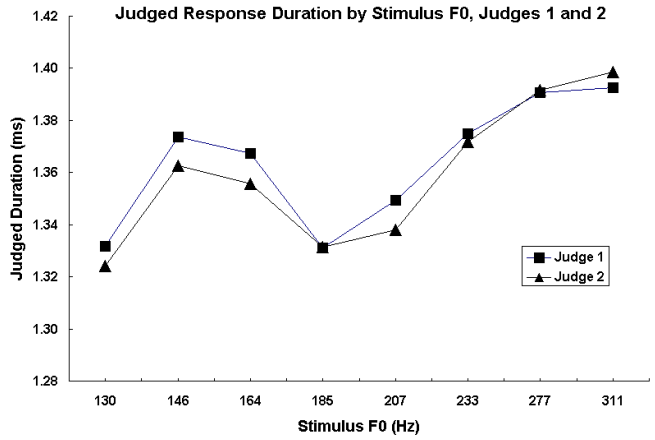

Figure 1- Judged response duration as a function of stimulus F0 for Judges 1 and 2

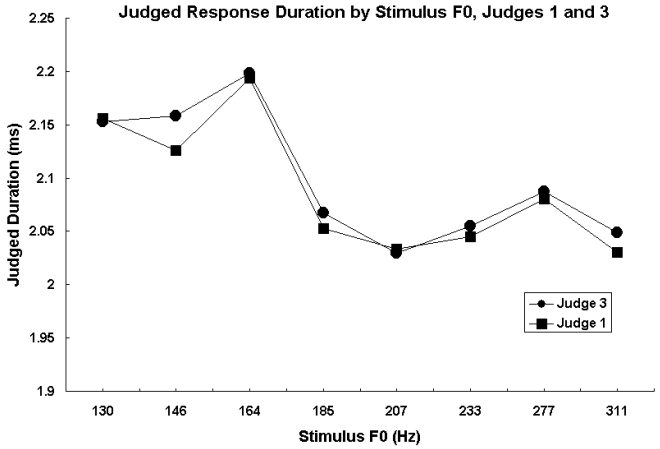

Figure 2- Judged response duration as a function of stimulus F0 for Judges 1 and 3

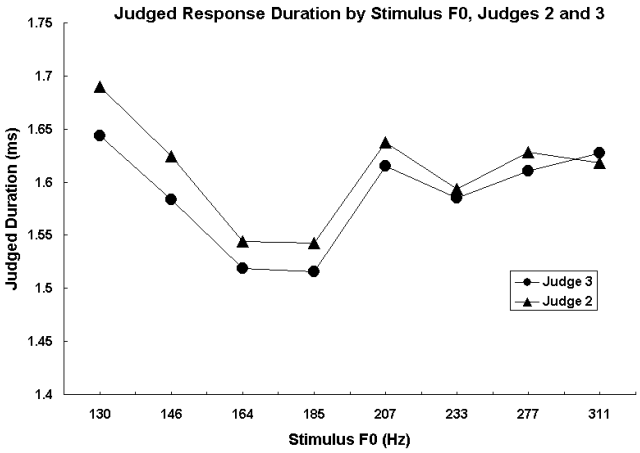

Figure 3- Judged response duration as a function of stimulus F0 for Judges 2 and 3

Figures 4-6 show extracted fundamental frequency values for pairs of judges as a function of stimulus fundamental frequency. As can be seen, reasonable agreement in F0 estimates is found for Judges 2 and 3 (Figure 6), but not for Judges 1 and 2 (Figure 4) or Judges 1 and 3 (Figure 5).

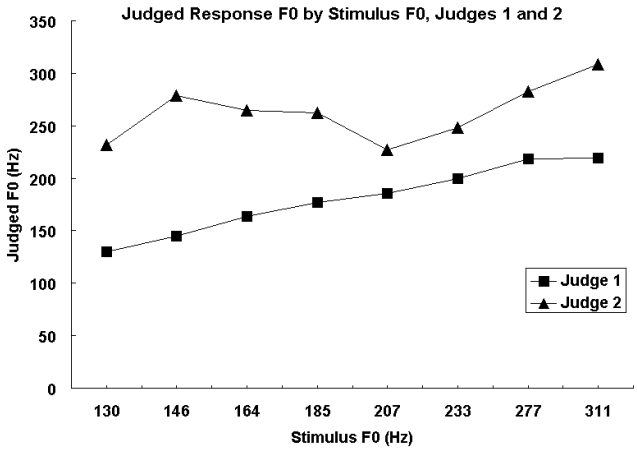

Figure 4- Judged response F0 as a function of stimulus F0 for Judges 1 and 2

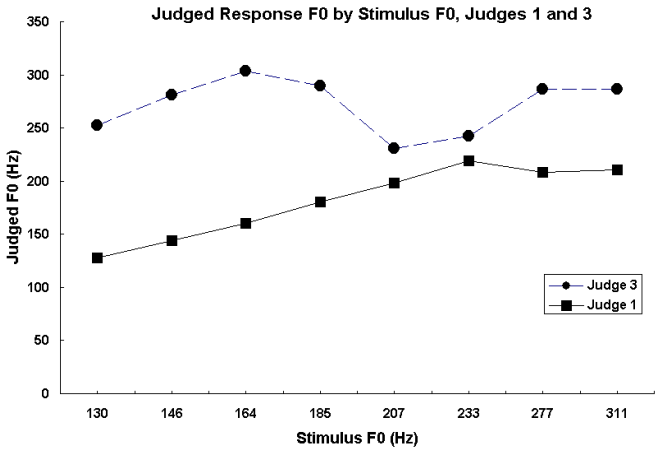

Figure 5- Judged response F0 as a function of stimulus F0 for Judges 1 and 3

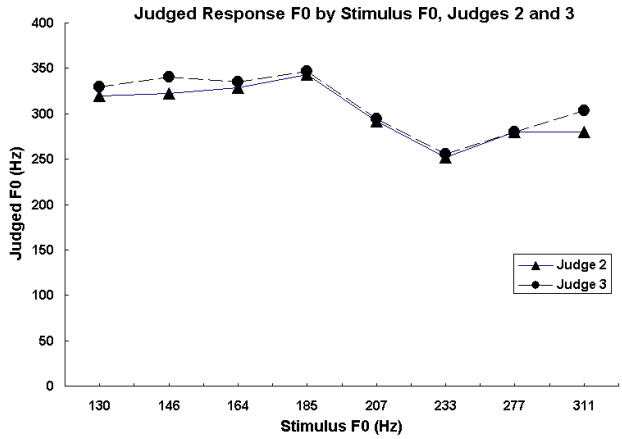

Figure 6- Judged response F0 as a function of stimulus F0 for Judges 2 and 3

Tables 1 and 2 show Pearson correlation coefficients for inter-rater reliability between pairs of judges. Table 1 reflects reliability in estimates of fundamental frequency; Table 2 reflects reliability in estimates of duration. The results show that the correlations among the three judges in terms of duration were very high. It is clear that the three judges could reliably identify the token and delineate its duration. However, the correlations among the judges in terms of fundamental frequency were very low, making this more sophisticated task unreliable.

	Judge 1	Judge 2	Judge 3
Judge 1	1.0		
Judge 2	0.24	1.0	
Judge 3	0.17	0.59	1.0

Table 1: Pearson correlations for inter-rater reliability in fundamental frequency estimates

	Judge 1	Judge 2	Judge 3
Judge 1	1.0		
Judge 2	0.99	1.0	
Judge 3	0.99	0.98	1.0

Table 2: Pearson correlations for inter-rater reliability in duration estimates

Chapter Four: Discussion and Conclusion

One possible explanation for the variability in fundamental frequency estimates arises from the fact that many of the samples, originally recorded in Cool Edit, were not sufficiently strong to be detected in the CSL program. CSL is a sophisticated program that looks for dysphonic voice problems rather than assuming that the waveform is periodic, which in this study it most often was. The judges in this study used the default settings of the software, and these settings may not have been appropriate for the task. The data, originally recorded on Cool Edit, was easily analyzed by that program. When transferring the data to CSL, there needed to be a boost in the amplitude in order for it to be read properly; however, the instrumentation to perform such a task was not available. Instead, the judges altered their methodology to accommodate for the low signal strength. Because no single set of strategies of accommodation was employed, each of the judges used slightly different methods to analyze the data. When the graph information was not readable, Judge 1 played the signal aloud to assist in identifying the highlighted token, and obtained the pitch contour. The graph of the pitch contour reflected the response in the window above, whether or not it could be detected visually. The response could then be selected in the top window by linking the two windows and selecting the response in the bottom window. Judge 2 used voiced period marks to select the response (see Masty, 2007, for further details). Judge 3 adjusted the window setting (see Finical, 2007, for further details).

In the previous studies, the judges were given the frequencies of the eliciting stimuli before making measurements of the response fundamental frequency. With this information the judges were able to easily determine the accuracy of their calculations. In this study such

information was not available to the judges; therefore, they were not able to immediately discern when estimates were highly inaccurate.

Topics for future research could include looking at the effect of specific instructions on interjudge reliability. Clear and uniform instructions would standardize the effect of different methodologies in the responses that are not easily discernable. It would also be beneficial to systematically evaluate the different strategies employed in analysis of the pitch matching to determine which one is most accurate. It is interesting to note that even though judges 2 and 3 used different strategies, their extracted pitches were quite similar. Overall, the present results suggest that the use of computer software packages in clinical voice situations can yield highly variable results unless users of the software are well-trained and given specific instructions for unusual measurement situations.

References

Ameer J. Time to phonatory stability in trained singers cued for pitch matching by pure-tone and synthesized voice. Unpublished Honors Thesis, (2003) The Ohio State University.

Curran K. Measurement of pitch matching accuracy: How much is too little? Unpublished Honors Thesis. (2004) The Ohio State University.

Ives S. The effects of sound duration and spectral complexity on pitch-matching accuracy in singers when cued with pure-tone and synthesized human voice models. Unpublished Master's Thesis, (2002) The Ohio State University.

Leonard R, Ringel R, Horri Y, Daniloff R. Vocal shadowing in singers and nonsingers. *J Speech Hear Res*. 1988; 31: 54-61.

Murry T. Pitch-matching accuracy in singers and nonsingers. *J Voice*. 1990; 4 (4): 317-21.

Slifken A, Newell K. Is variability in human performance a reflection of system noise. *Current Directions in Psychological Science*. 1998; 7 (6): 170-77.

Small AR, McCachern FL. The effects of male and female vocal modeling on pitch-matching accuracy of first-grade children. *JRME*. 1983; 31: 227-33.

Smits I, Ceuppens P, De Bodt M. A comparative study of acoustic voice measurements by means of Dr. Speech and Computerized Speech Lab. *J Voice*. 2005; 19 (2):187-96.

Tervaniemi et al. Harmonic partials facilitate pitch discrimination in humans: electrophysiological and behavioral evidence. *Neuroscience Letters*. 2000; 279: 29-32.

Yarbrough C, Bower J, Benson W. The effects of vibrato on pitch-matching accuracy of certain and uncertain singers. *JRME*. 1992; 40 (1): 30-38.